BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 8: e135729
doi: 10.3897/biss.8.135729

OPEN ACCESS

Conference Abstract

METAPLANTCODE: Harmonizing Plant Metabarcoding Pipelines in Europe

Auguste Gardette‡, Youcef Sklab‡, Eugeni Belda‡, Eric Chenin‡, Jean-Daniel Zucker‡

‡ IRD, Sorbonne Université, UMMISCO, F-93143, Bondy, France

Corresponding author: Auguste Gardette (auguste.gardette@ird.fr), Youcef Sklab (youcef.sklab@ird.fr), Eugeni Belda (eugeni.belda@ird.fr)

Received: 28 Aug 2024 | Published: 28 Aug 2024

Citation: Gardette A, Sklab Y, Belda E, Chenin E, Zucker J-D (2024) METAPLANTCODE: Harmonizing Plant Metabarcoding Pipelines in Europe. Biodiversity Information Science and Standards 8: e135729. https://doi.org/10.3897/biss.8.135729

Abstract

The METAPLANTCODE project is dedicated to advancing and optimizing pan-European case studies on metabarcoding. The project's objectives include providing best practice recommendations, optimizing analysis pipelines for species identification, and creating user-friendly reference databases. To accomplish these objectives, METAPLANTCODE will identify and address gaps in current methodologies, publish best practice documents on FAIR (Findable, Accessible, Interoperable, Reusable) data publishing for plant metabarcode data to GBIF (Global Biodiversity Information Facility) and the INSDC (International Nucleotide Sequence Database Collaboration), and implement ELIXIR-compatible multimodal deep learning (DL) models in novel tools for standalone metabarcoding analyses using various data sources.

A significant focus of the project is enhancing species identification accuracy through GBIF records and metadata. This involves mapping regional, national, and international botanical taxonomic checklists, red lists, and floras to the Catalogue of Life (COL) via the COL ChecklistBank. Additionally, taxonomic and floristic literature will be semantically enriched with new entity recognition and relationship extraction modules, supporting the enhanced identification of species through domain-specific descriptive and phenotypic features. An interface will link taxonomic names to treatments, identify homonyms and synonyms, and facilitate the conversion and annotation of floras, red lists, and ecological

© Gardette A et al. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

treatments. All METAPLANTCODE products will adhere to FAIR standards by the project's end.

The project emphasizes knowledge transfer from the outset, engaging with associated partners and stakeholders. Key stakeholders will be identified, priorities set, and communication channels established, monitored, and adjusted as necessary. Efforts to enhance stakeholder engagement, training, and outreach will ensure that plant metabarcoding becomes a routine standard for biodiversity monitoring in Europe and beyond.

Deep Learning for Plant Metabarcoding

Within the METAPLANTCODE project, our team is tasked with improving taxonomic precision by integrating deep learning on metabarcoding data and metadata. Previous studies have demonstrated the applicability of deep learning to non-plant barcoding data and its computational efficiency compared to traditional bioinformatics approaches (Flück et al. 2022).

Deep Learning Models for Metabarcoding Data

Our approach involves evaluating the efficacy of several deep learning models—such as Convolutional Neural Networks (CNN)(LeCun et al. 2015), Transformer models (Vaswani et al. 2017), Hyena (Poli et al. 2023), and Mamba architectures (Gu et al. 2023)—on plant barcoding datasets. Preliminary results will be presented, highlighting the application of these models and the proposed ensemble method (Mohammed and Kora 2023), which combines multiple barcode sequence representations and learning strategies. The ensemble approach, when integrated with classical machine learning models such as logistic regression and Support Vector Machines (SVM) (Noble 2006), is anticipated to offer improved precision and robustness compared to individual model applications (Fig. 1).

Models	Accuracy	F1 Score	Recall	Precision
Logistic Regression	0.938	0.939	0.939	0.949
SVM	0.941	0.941	0.940	0.950
DNABERT	0.904	0.839	0.840	0.845
DNABERT 2	0.535	0.386	0.398	0.394
DNABERT S	0.825	0.731	0.739	0.733
AgroNucleotide Transformer	0.893	0.816	0.824	0.817
Nucleotide Transformer	0.917	0.868	0.873	0.871
HyenaDNA	0.821	0.718	0.723	0.723
GNN	0.843	0.840	0.843	0.876
Ensemble Method	**0.987**	**0.986**	**0.987**	**0.989**

Figure 1.

Performance comparison of machine learning and deep learning models on a dataset of 156 plant species using 16 barcodes per species from PLANiTS database (Banchi et al. 2020).

Multimodal Refinement of Predictions

In the subsequent phase, we aim to refine genetic sequence classifications by employing a multimodal strategy. This approach will integrate genetic information with traditional botanical knowledge. We will utilize biological interaction lists (e.g., species-species, species-habitat) provided by the METAPLANTCODE project to train a large language model (LLM) on relevant scientific literature. This LLM, specifically tailored for plant biodiversity, will incorporate metadata associated with genetic samples (including location, temporality, and climatic conditions). By merging embeddings of both metadata and genetic data, we aim to enhance the accuracy of taxonomic predictions (Fig. 2).

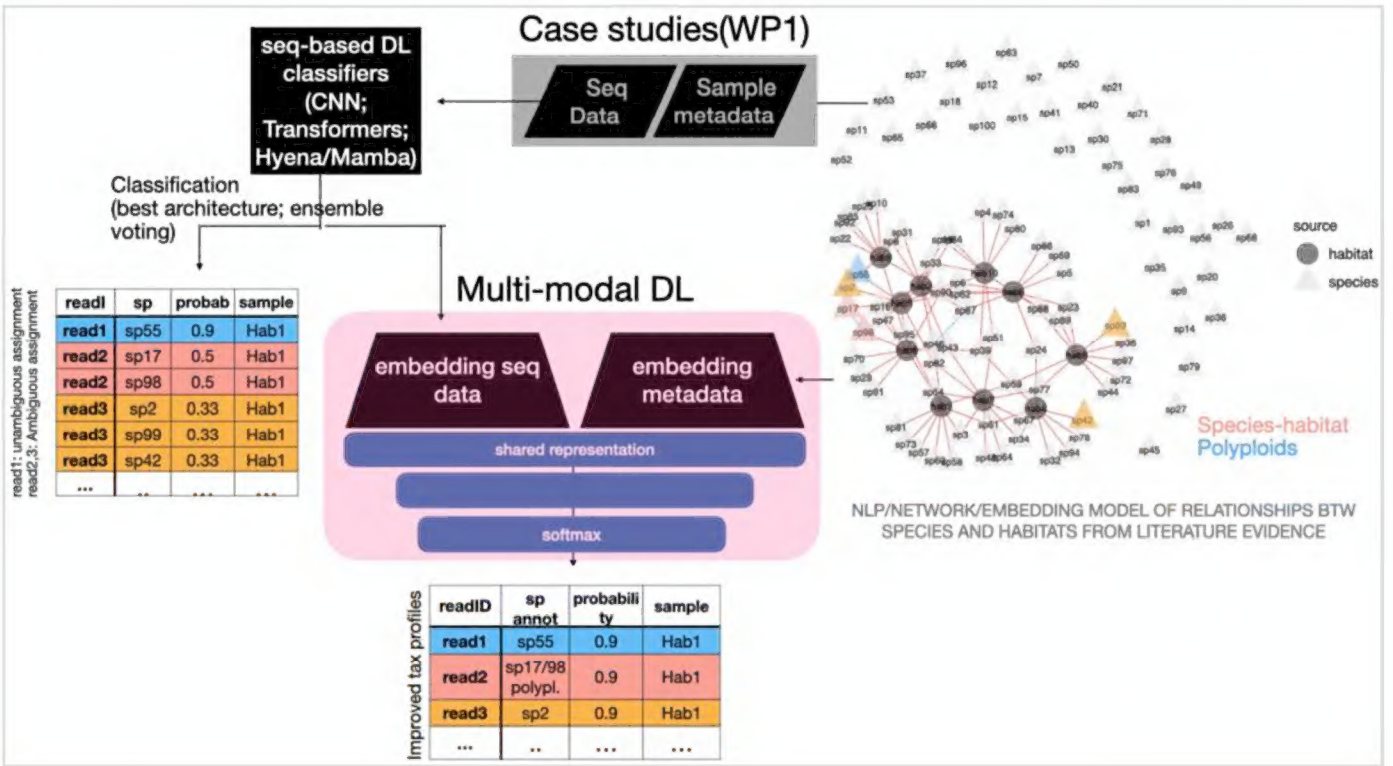
Case studies(WP1)
seq-based DL classifiers (CNN; Transformers; Hyena/Mamba)
Seq Data
Sample metadata
Classification (best architecture; ensemble voting)
read1: unambiguous assignment
read2,3: Ambiguous assignment
readI | sp | probab | sample
read1 | sp55 | 0.9 | Hab1
read2 | sp17 | 0.5 | Hab1
read2 | sp98 | 0.5 | Hab1
read3 | sp2 | 0.33 | Hab1
read3 | sp99 | 0.33 | Hab1
read3 | sp42 | 0.33 | Hab1
Multi-modal DL
embedding seq data
embedding metadata
shared representation
softmax
source
habitat
species
Species-habitat
Polyploids
NLP/NETWORK/EMBEDDING MODEL OF RELATIONSHIPS BTW SPECIES AND HABITATS FROM LITERATURE EVIDENCE
Improved tax profiles
readID | sp annot | probability | sample
read1 | sp55 | 0.9 | Hab1
read2 | sp17/98 polypl. | 0.9 | Hab1
read3 | sp2 | 0.9 | Hab1

Figure 2.

Proposed multimodal integration framework for enhancing taxonomic predictions in plant metabarcoding.

- **Upper Left**: The genomics module represents models that generate initial taxonomic predictions based on sequencing data.
- **Upper Right**: A graph illustrates the biodiversity knowledge, showcasing relationships between species and habitats. This knowledge is derived from biological interaction lists (e.g., species-species, species-habitat).
- **Center**: The framework's core is a multimodal approach using a Large Language Model (LLM) trained on plant biodiversity literature. This LLM integrates metadata (e.g., location, temporality, climate) with genetic data, combining their embeddings to improve taxonomic predictions.

Conclusion

Through this research, we aim to develop an effective method for integrating genetic data with textual information from various sources. We anticipate that this approach will not only enhance plant metabarcoding but also be applicable to other barcoding fields, such

as bacteria, fish, fungi, and more. Additionally, we expect this methodology to find broader applications in genomic research, providing valuable insights and improvements across diverse biological disciplines.

Keywords

AI, multi-modality, biodiversity, barcoding, botany, herbarium

Presenting author

Auguste Gardette

Presented at

SPNHC-TDWG 2024

Funding program

Biodiversita+ is a European Biodiversity Partnership supporting excellent research on biodiversity with an impact for society and policy.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Banchi E, Ametrano CG, Greco S, Stankovic D, Muggia L, Pallavicini A (2020) PLANiTS: a curated sequence reference dataset for plant ITS DNA metabarcoding. Database 2020 https://doi.org/10.1093/database/baz155
- Flück B, Mathon L, Manel S, Valentini A, Dejean T, Albouy C, Mouillot D, Thuiller W, Murienne J, Brosse S, Pellissier L (2022) Applying convolutional neural networks to speed up environmental DNA annotation in a highly diverse ecosystem. Scientific Reports 12 (1). https://doi.org/10.1038/s41598-022-13412-w
- Gu, Albert, Dao, Tri (2023) Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.0075.
- LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521 (7553): 436-444. https://doi.org/10.1038/nature14539
- Mohammed A, Kora R (2023) A comprehensive review on ensemble deep learning: Opportunities and challenges. Journal of King Saud University - Computer and Information Sciences 35 (2): 757-774. https://doi.org/10.1016/j.jksuci.2023.01.014

- Noble WS (2006) What is a support vector machine? Nature Biotechnology 24 (12): 1565-1567. https://doi.org/10.1038/nbt1206-1565
- Poli M, Massaroli S, Nguyen E, Fu D, Dao T, Baccus S, Bengio Y, Ermon S, Ré C (2023) Hyena Hierarchy: Towards Larger Convolutional Language Models. Proceedings of the 40th International Conference on Machine Learning. Honolulu, Hawaii, USA. 35 pp. URL: https://proceedings.mlr.press/v202/poli23a/poli23a.pdf
- Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, Polosukhin (2017) Attention is all you need. Advances in Neural Information Processing Systems https://doi.org/10.48550/arXiv.1706.03762